

Learning Residual Images for Face Attribute Manipulation

Wei Shen Rujie Liu

Fujitsu Research & Development Center, Beijing, China.

{shenwei, rjliu}@cn.fujitsu.com

Abstract

Face attributes are interesting due to their detailed description of human faces. Unlike prior research working on attribute prediction, we address an inverse and more challenging problem called face attribute manipulation which aims at modifying a face image according to a given attribute value. In order to obtain an efficient representation for the manipulation, we propose to learn the corresponding residual image which is defined as the difference between images after and before the manipulation. Using the residual image, the manipulation can be operated efficiently with modest pixel modification. The framework of our approach is based on the Generative Adversarial Network. It consists of two image transformation networks imitating the attribute manipulation and its dual operation and a shared discriminative network distinguishing the generated images from different reference images. We also apply dual learning to allow the two transformation networks to learn from each other. Experiments show that the learned residual images successfully simulate the manipulations and the generated images retain most of the details in attribute-irrelevant areas.

1. Introduction

Considerable progresses have been made on face image processing, such as age analysis [23][29], emotion detection [1][5] and attribute classification [4][20][15][18]. Most of these studies concentrate on inferring attributes from images. However, we raise an inverse question on whether we can manipulate face images towards a desired attribute with modest modification (i.e. face attribute manipulation). Some examples are shown in Fig. 1. Since the modified images have almost the same identity information, they can potentially serve as augmented data or preprocessed data for some specific applications.

Face attribute manipulation can be regarded as a process of image generation from the latent space in which face attributes embedded in. Modifying a face attribute in pixel space is equivalent to moving its corresponding latent rep-

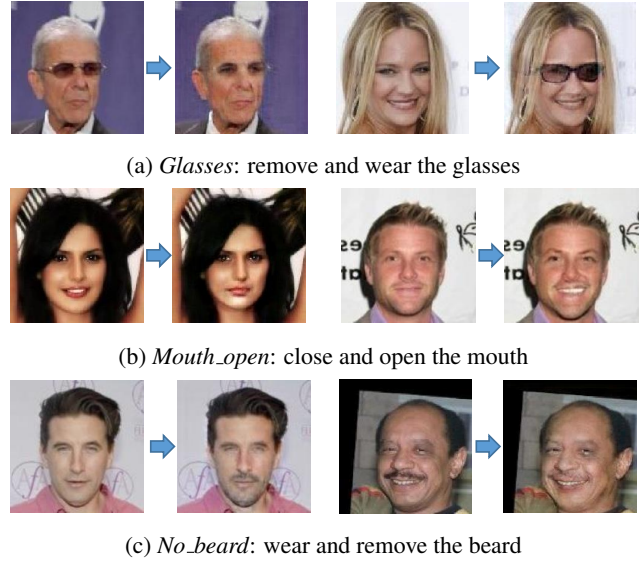


Figure 1: Illustration of face attribute manipulation. From top to bottom are the manipulations of *glasses*, *mouth_open* and *no_beard*.

resentation from its original point to the target point which is used for reconstruction. Recently, generative models such as the generative adversarial network (GAN) [7] and the variational autoencoder (VAE) [14] emerged as powerful models capable of generating images from latent representations. Images generated from the GAN model are sharp and realistic. However, the model can not encode images since it is the random noise that the model uses for image generation. Compared to the GAN model, the VAE model is able to encode the given image to a latent representation. Nevertheless, passing images through the encoder-decoder pipeline often harms the quality of the reconstruction. In the scenario of face attribute manipulation, those details can be identity-related and the loss of those details will cause undesired visually drastic change. Thus, it is difficult to directly employ the GAN model or the VAE model to face attribute manipulation.

An alternative way is to view face attribute manipulation as a transformation process which takes in original images as input and then outputs transformed images without explicit embedding. Such a transformation process can be efficiently implemented by a feed-forward convolutional neural network (CNN). When manipulating face attributes, the feed-forward networks are required to modify the attribute-specific area and keep irrelevant areas unchanged both of which are challenging.

In this paper, we propose a novel method based on residual image learning for face attribute manipulation. The method combines the generative power of the GAN model with the efficiency of the feed-forward network (see Fig. 2). As is demonstrated that learning residual functions with reference to the layer inputs leads to easy optimization and high classification accuracy [9], we also provide evidence that residual learning is effective in face attribute manipulation. We model the manipulation operation as learning the residual image which is defined as the difference between the desired manipulated image and the original input image. Compared to learning to generate the manipulated image, learning the residual image is an easier task since it does not need to reconstruct most details of the face image and it focuses on the attribute-specific part of the face. In other words, it avoids learning redundant information by concentrating on the essential attribute-specific knowledge. To improve the efficiency of manipulation learning, we adopt two CNNs to model two inverse manipulations (*e.g.* removing glasses as the primal manipulation and wearing glasses as the dual manipulation, see Fig. 2) and apply the strategy of dual learning during the training phase. Our contribution can be summarized as follows.

1. We propose to learn residual images for face attribute manipulation. The proposed method simplifies the manipulation task by manipulating only the attribute-specific face area instead of reconstructing the entire face which involves redundant computation of irrelevant details.
2. We devise an integrated scheme to learn two inverse attribute manipulations simultaneously with a shared discriminator. The formation of this scheme enables us to model one manipulation task as the primal task and the other as the dual task and helps us gain benefits from the strategy of dual learning.
3. We show that wearing glasses can affect the face landmark detection algorithm. However, this harm can be alleviated via applying glasses removal using the proposed method.

2. Related Work

There are many techniques for image generation in recent years [24][2][17][8][3][14]. Most of them fall into

the category of unsupervised learning. Radford *et al.* [24] applied deep convolutional generative adversarial networks (DCGANs) to learn a hierarchy of representations from object parts to scenes for general image generation. Chen *et al.* [2] introduced an information-theoretic extension to the GAN that is able to learn disentangled representations which can be used to generate new images. Larsen *et al.* [17] combined the VAE with the GAN to learn an embedding in which high-level abstract visual features can be modified using simple arithmetic.

Although unsupervised learning can disentangle latent attribute representation. Reconstructions from the direct arithmetic of those representations may cause undesired problems, *e.g.* highly correlated visual features [17]. Our work is an independent work along with [19]. In [19], Li *et al.* proposed a deep convolutional network model for identity-aware transfer of facial attributes. The differences between our work and [19] are noticeable in three aspects. (1) Our method generates the manipulated image using the residual image which is different from [19]. (2) Our method models two inverse manipulations within a single architecture while the work in [19] treats each manipulation independently. (3) Our method does not need post-processing which is essential in [19].

3. Learning the Residual Image

The architecture of the proposed method is presented in Fig. 2. For each face attribute manipulation, it contains two image transformation networks G_0 and G_1 and a discriminative network D . G_0 and G_1 simulate the primal and the dual manipulation respectively. D classifies the reference images and generated images into three categories. The following sections will first give a brief introduction of the generative adversarial network and then the detailed description of the proposed method.

3.1. Generative Adversarial Networks

The generative adversarial network is introduced by Goodfellow *et al.* [7]. It is an unsupervised framework containing a generative model G and a discriminative model D . The two models play a minimax two-player game in which G tries to recover the training data and fool D to make a mistake about whether the data is from realistic data distribution or from G . Given a data prior p_{data} on data x and a prior on the input noise variables $p_z(z)$, the formal expression of the minimax game is as follows:

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]. \quad (1)$$

The parameters of both G and D are updated iteratively during the training process. The GAN model provides an

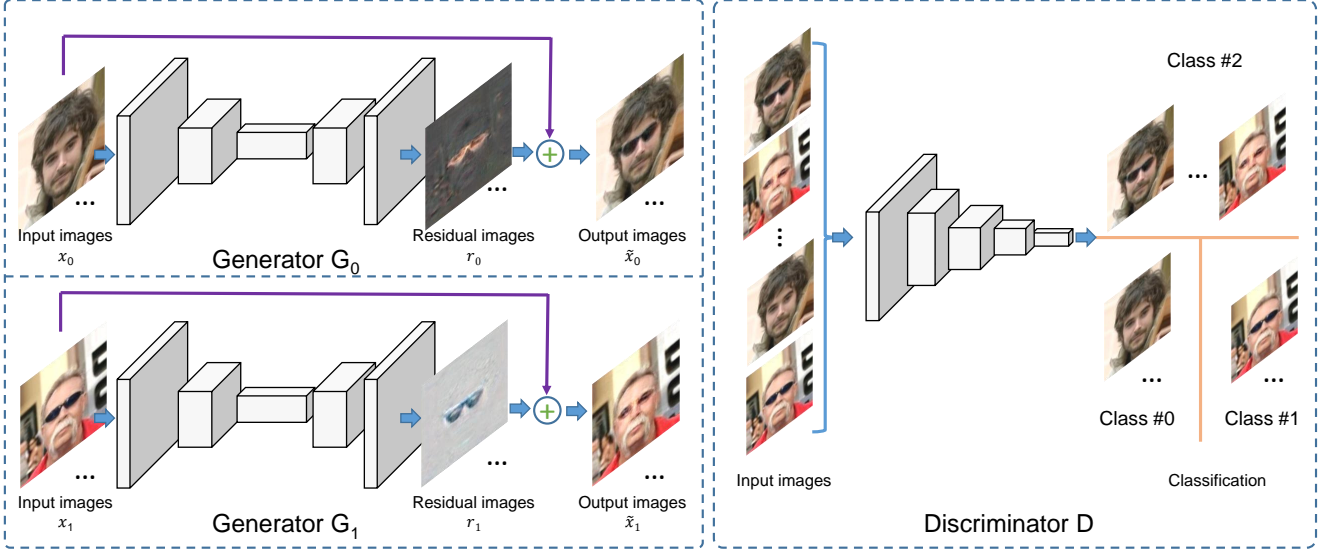


Figure 2: The architecture of the proposed method. Two image transformation networks G_0 and G_1 perform the inverse attribute manipulation (*i.e.* wear glasses and remove glasses). Both G_0 and G_1 produce the residual images with reference to the input images. The final output images are the pixel-wise addition of the residual images and the input images. The discriminative network D is a three-category classifier that classifies images from different distribution (*i.e.* images generated from G_0 and G_1 , images with positive attribute labels and images with negative attribute labels).

effective way to learn data distribution of realistic images and makes it possible to generate images with desired attribute. Based on the GAN model, we redesign the generator and the discriminator for face attribute manipulation in the following sections.

3.2. Image Transformation Networks

The motivation of our approach is that face attribute manipulation usually needs modest modification of the attribute-specific face area while other parts remain unchanged. For example, when removing a pair of glasses from a face image, only the area of the glasses should be replaced with face skin or eyes and the change of other face parts such as mouth, nose, and hair should not be involved. Thus, we model the manipulation as learning the residual image targeted on the attribute-specific area.

As shown in Fig. 2, G_0 and G_1 are used to simulate the manipulation and its dual operation. Given the input face image x_0 with a negative attribute value and the input face image x_1 with a positive attribute value, the learned network G_0 and G_1 apply the manipulation transformations to yield the residual images r_0 and r_1 . Then the input images are added to the residual images as the final output of the manipulation result \tilde{x}_0 and \tilde{x}_1 :

$$\tilde{x}_i = x_i + r_i = x_i + G_i(x_i), i = 0, 1. \quad (2)$$

In order to let the residual image be sparse, we apply an

L-1 norm regularization as

$$\ell_{pix}(r_i) = \|r_i\|_1. \quad (3)$$

The image transformation networks differ from the generative CNN of the vanilla GAN [7] in the input. The input of our image transformation networks is the face image while that of the GAN is some random noise. Our transformation networks also differ from convolutional auto-encoder [22] since it does not try to reconstruct the input image.

3.3. The Discriminative Network

Given the real images x_0 and x_1 with known attribute label 0 and label 1, we regard the transformed images \tilde{x}_0 and \tilde{x}_1 as an extra category with label 2. The loss function is the softmax loss:

$$\ell_{cls}(t, p) = -\log(p_t), t = 0, 1, 2, \quad (4)$$

where t is the label of the image and p_t is the softmax probability of the t -th label. Similar strategy for constructing the GAN loss is also adopted in [27]. Given the discriminative model D , the loss function for the image transformation networks G_0 and G_1 is

$$\ell_{GAN} = \begin{cases} -\log(D(G_i(x_i))) & i = 0, \\ -\log(1 - D(G_i(x_i))) & i = 1. \end{cases} \quad (5)$$

Image transformation networks G_0/G_1	Discriminator D
Input 128×128 color images	Input 128×128 color images
5×5 conv. 64 leaky RELU. stride 1. batchnorm	4×4 conv. 64 leaky RELU. stride 2. batchnorm
4×4 conv. 128 leaky RELU. stride 2. batchnorm	4×4 conv. 128 leaky RELU. stride 2. batchnorm
4×4 conv. 256 leaky RELU. stride 2. batchnorm	4×4 conv. 256 leaky RELU. stride 2. batchnorm
3×3 conv. 128 leaky RELU. stride 1. upsampling. batchnorm	4×4 conv. 512 leaky RELU. stride 2. batchnorm
3×3 conv. 64 leaky RELU. stride 1. upsampling. batchnorm	4×4 conv. 1024 leaky RELU. stride 2. batchnorm
4×4 conv. 3	4×4 conv. 1

Table 1: The network architecture of the image transformation networks G_0/G_1 and the discriminator D

Perceptual loss is widely used to measure the content difference between different images [11][6][17]. We also apply this loss to encourage the transformed image to have similar content to the input face image. Let $\phi(x)$ be the activation of the third layer in D . The perceptual loss is defined as:

$$\ell_{per}(x, \tilde{x}) = \|\phi(x) - \phi(\tilde{x})\|_1. \quad (6)$$

3.4. Dual Learning

In addition to applying adversarial learning in the model training, we also adopt dual learning which has been successfully applied in machine translation [30]. A brief introduction is as follows. Any machine translation has a dual task, *i.e.* source language to target language (primal) and target language to source language (dual). The mechanism of dual learning for machine translation can be viewed as a two-player communication game. The first player translates a message from language A to language B and sends it to the second player. The second player checks if it is natural in language B and notifies the first player. Then he translates the message to language A and sends back to the first player. The first player checks whether the received message is consistent with his original message and notifies the second player. The information feedback signals from both players can benefit each other through a closed loop.

The dual learning process in this work is implemented as Fig. 3. For a given image x_0 with a negative attribute value, we pass it through the transformation network G_0 . The obtained image $\tilde{x}_0 = G_0(x_0)$ is then fed to the transformation network G_1 . The yielded image is $\hat{x} = G_1(\tilde{x}_0) = G_1(G_0(x_0))$. Since G_0 and G_1 are the primal task and the dual task respectively, \hat{x} is expected to have the same attribute value as x_0 . Similar process is also applied for x_1 . The loss function for the transformation networks is expressed as:

$$\ell_{dual}(\tilde{x}_i) = \begin{cases} -\log(1 - D(G_{1-i}(\tilde{x}_i))) & i = 0, \\ -\log(D(G_{1-i}(\tilde{x}_i))) & i = 1. \end{cases} \quad (7)$$



Figure 3: The dual learning process in this work.

3.5. Loss Function

Taking the loss functions all together, we have the following loss function for G_0/G_1 :

$$\ell_G = \ell_{GAN} + \ell_{dual} + \alpha \ell_{pix} + \beta \ell_{per}, \quad (8)$$

where α and β are constant weight for regularization terms.

For D , the loss function is

$$\ell_D = \ell_{cls}. \quad (9)$$

In this work, we keep $\beta = 0.1\alpha$ and set $\alpha=5e-4$ for local face attribute manipulation and $\alpha=1e-6$ for global face attribute manipulation.

4. Experiments

As current evaluation methods (*e.g.* average log-likelihood and Parzen window estimates) are problematic for measuring the quality of generative models [28]. We firstly qualitative compare the results from different models and then quantify the effectiveness of the proposed method in glasses removal in Sec. 4.5. The detailed architectures of G_0 , G_1 and D are specified in Tab. 1.

4.1. Datasets

We select two datasets in our experiments, *i.e.* the CelebA dataset and the Labeled Faces in the Wild (LFW) dataset [10] with attribute annotations from [15]. The CelebA dataset [21] contains more than 200K celebrity images, each with 40 binary attributes. We picked 6 of them,

i.e. glasses, mouth_open, smile, no_beard, young, and male to evaluate the proposed method. The center part of the aligned images in the CelebA dataset are cropped and scaled to 128×128 . Despite there are a large number of images in the dataset, the attribute labels are highly biased. Thus, for each attribute, 1,000 images from the attribute-positive class and 1,000 images from the attribute-negative class are randomly selected for test. From the rest images, we select all the images belong to the minority class and equal number of images from the majority class to make a balance dataset. The LFW dataset is used only for testing the generalization of our method. Note that there are no ground truth manipulated images for training the transformation networks since the images in the CelebA are all obtained from the Internet.

4.2. Local and Global Attribute Manipulation

Among the six attributes, we group *glasses*, *mouth_open*, *smile* and *no_beard* as local attributes since the manipulation only operates on local face area. The other two attributes *male* and *young* are global attributes. We compare the results of our method and that of the state-of-the-art method VAE-GAN [17] on the CelebA dataset (Fig. 4). The results on the LFW dataset are presented in Fig. 5.

We firstly give an overall observation of the results. As shown in Fig. 4, the VAE-GAN model [17] changes many details, such as hair style, face color and background objects. In contrast, the results from our method retains most of the details. Comparing the original images in the first row and the transformed images in the third row, we can find that the details of the original faces mostly remain except the area corresponding to the manipulated attribute. This is also proved by the residual images in the last row. For local attribute manipulation, the strong responses on the residual images mainly concentrate in local areas. For example, when adding sun glasses to the face image, the most strong response on the residual image is the black sun glasses. Similarly, removing glasses will cause the residual image to enhance the eyes and remove any hint of glasses that is presented in the original face image.

Local face attribute manipulations are straightforward and obvious to notice. Since the manipulations of *glasses* and *no_beard* only require simple operations like removal or filling in, we investigate more interesting tasks such as *mouth_open* and *smile* manipulations. Both manipulations will cause the “movement” of the chin. From Fig. 4(c,d), we can observe that the “movement” of the chin is captured by the image transformation networks. When performing the *mouth_open* manipulation, the network “lowers” the chin and when performing *mouth_close* manipulation, the network “lifts” the chin.

The most challenging task would be manipulating global attributes *young* and *male*. The networks have to learn subtle differences such as wrinkles, hair color, beard *etc.* In



Figure 6: Validation of residual image learning and dual learning in face attribute manipulation. First row: the original input images. Second row: the result images from the proposed model. Third row: the result images from the model without residual image learning. Last row: the result images from the model without dual learning.

Fig. 4(e,f), changing from young to old will cause more wrinkles and the reverse operation will also darken hair color. The main difference between the male and the female are the beard, the lipstick and the eyes. The strong responses in the residual images for these two manipulations are scattered over the entire images rather than restricted within a local area.

4.3. Ablation Study

Our model consists of two pivotal components: residual image learning and dual learning. In this section, we further validate their effectiveness. The validation of residual image learning is performed by blocking the identity mapping in the transformation network which enforces the network to learn to generate the entire image. The validation of dual learning is implemented by breaking the data-feed loop, *i.e.* the output of G_0 and G_1 will not be fed to each other. Other network settings are kept the same as the proposed model. The results are shown in Fig. 6. We observe that without residual image learning, the model produces much lower quality images in terms of introducing much noise and some wrongly added beard in the second and third column which indicates the task has become challenging. The drop of dual learning also deteriorates image quality. We notice some change in hair color which may be caused by the performance degradation of G_0 , G_1 and D . Thus, we argue that combining residual image learning with dual learning will lead to better attribute manipulation results.

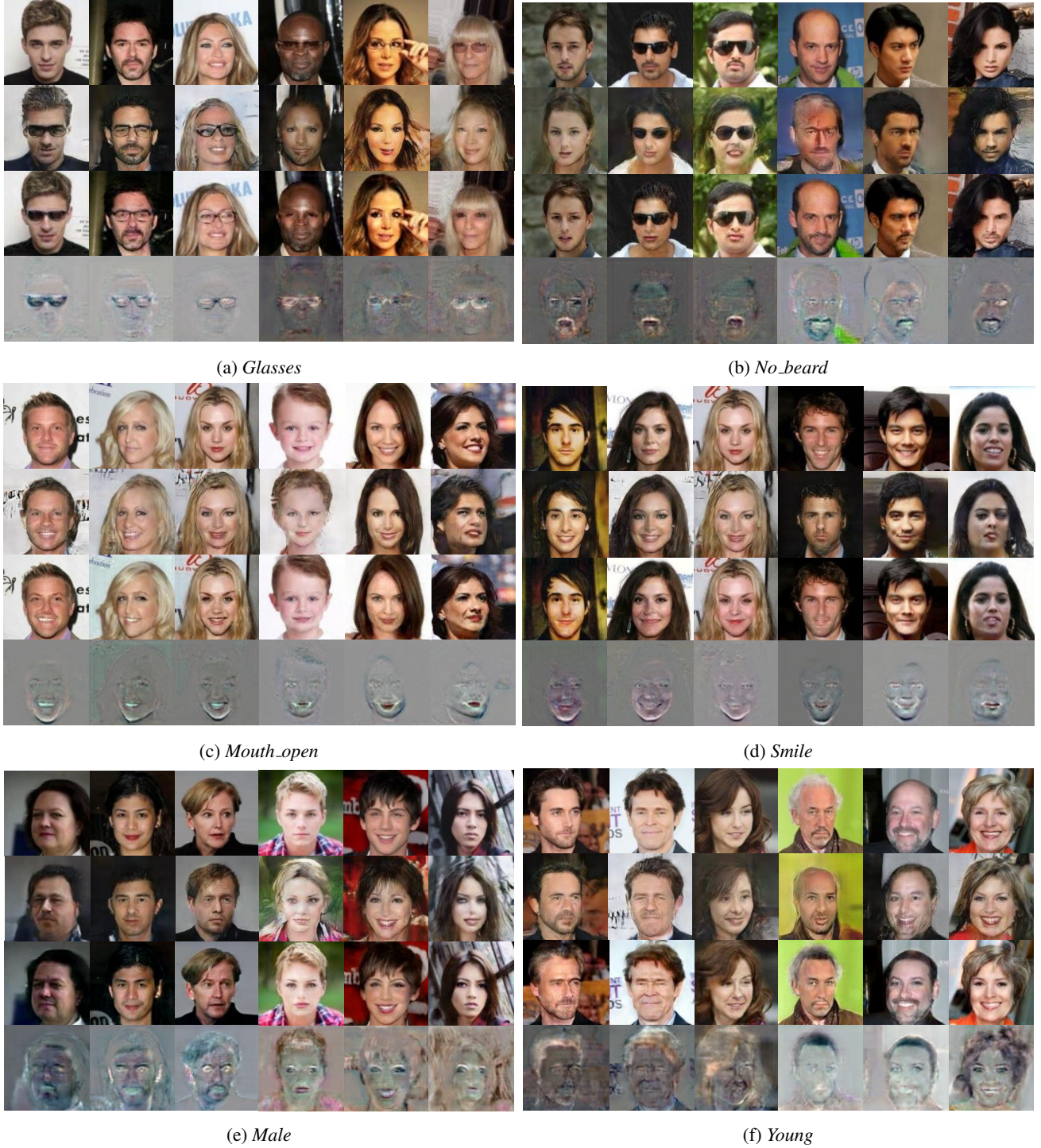


Figure 4: Face attribute manipulation on the CelebA dataset. For each sub-figure, the first row are the original face images. The second and the third row are the generated images with opposite attribute values using the VAE-GAN method [17] and the proposed method respectively. The last row are the residual images learned by the proposed method. Results of the primal task and the dual task of each attribute manipulation are presented in the first three columns and the last three columns respectively.

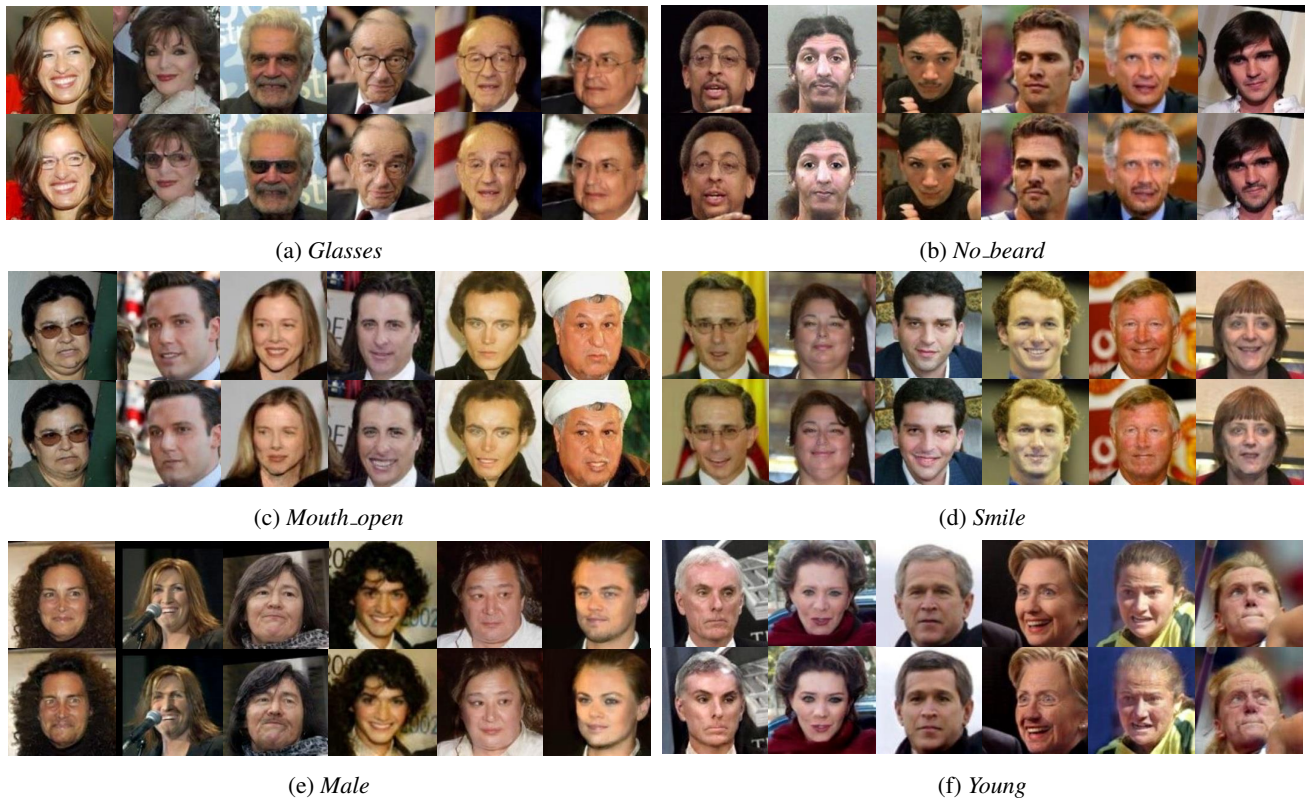


Figure 5: Face attribute manipulation on the LFW dataset. For each sub-figure, results of the primal task and the dual task of each attribute manipulation are presented in the first three columns and the last three columns respectively.



Figure 7: Visual feature decorrelation in the manipulation of attribute *no_beard*. First row: the original input images. Second row: the reconstructed images from the VAE-GAN model [17]. Third row: the manipulated images from the VAE-GAN model [17]. Last row: the resulted images from the proposed model.

4.4. Visual Feature Decorrelation

Training a classifier in an end-to-end way does not ensure the classifier can precisely identify the target visual features especially when the dataset is highly biased. Blind spots of predictive models are observed in [16]. For example, if the training data consists only of black dog images, and white cat images. A predictive model trained on these data will incorrectly label a white dog as a cat with high confidence in the test phase. Fortunately, this extremely biased distribution is not always the common case if more data can be collected. When analyzing the CelebA dataset, we find that *male* and *no_beard* is highly correlated (the Pearson correlation is -0.5222). It is not surprising since only the male can wear beard.

Classifiers trained with correlated features may also propagate the blind spots back to the generative models causing the generators to produce correlated visual features. To demonstrate the proposed method can learn less correlated features, we choose to manipulate the attribute *no_beard* on the female. It is extremely difficult since there is no any image containing a female face wearing beard in the training data. Adding beard to a female face may in-

roduce other male features. We compare the manipulation results from our method with those from the VAE-GAN model [17] in Fig. 7. We show the VAE-GAN reconstruction results of the original images to ensure that the VAE-GAN model do learn well about the original images. However, it is obvious that the VAE-GAN model changes many details. The hair length in the manipulated images is significantly shorter than that in the original images. This could be explained by the fact that most men wear short hair and this visual feature is correlated with the beard feature. The VAE-GAN model incorrectly treats the short-hair feature as the evidence of the presence of the beard. However, the proposed method successfully decorrelates these two features. The hair in the transformed images is the same as that in the original images. We owe this appealing property to residual image learning and dual learning that help the method concentrate on attribute-specific area.

4.5. Landmark Detection with Glasses Removal

Besides visually inspecting the manipulation results, we quantify the effectiveness of glasses removal using the proposed method in face landmark detection. The landmark detection algorithm is based on [12] and is implemented in *Dlib* [13]. We trained the detection model on the 300-W dataset [25][26]. Three test sets are included into consideration. They are dataset $D1$ containing images wearing glasses, dataset $D0$ containing images without glasses and dataset $D1_m$ containing the same images as $D1$ except that those images are processed with glasses removal. Note that $D0$ and $D1$ are the same test sets as that in the experiment of *glasses* manipulation.

The detection results are reported in Tab. 2 and illustrated in Fig. 8. Although the identities in $D0$ and $D1$ are different, we can still find that wearing glasses effects the landmark detection. Comparing the results of the eye landmarks detection between the first and the second column, we find the detection error increases. However, the error on $D1_m$ is much lower than that on $D1$ which demonstrates the benefit of applying glasses removal for eye landmark detection. Comparing the second row in Tab. 2, we observe that the errors on $D1$ and $D1_m$ are almost the same which indicates that the rest parts of the face remain almost unchanged.

5. Discussion

The hyper-parameters α and β are manually set in this work. Since manipulating global face attributes requires modifying larger area of the image than that is required by manipulating local face attributes, the hyper-parameters should be much smaller to allow the global modification. We empirically found $\alpha=5e-6$ and $\alpha=5e-4$ are appropriate for global and local attribute manipulation respectively. Larger α and β will prevent the generative models from simulating the manipulations and will collapse the models



Figure 8: Performance improvement on landmark detection brought by glasses removal. The ground truth landmark points and the detected landmark points on $D1$ are shown as green points and red points respectively. The detected landmark points after glasses removal using the proposed method are shown as yellow points.

Landmark	$D0$	$D1$	$D1_m$
Eye landmarks	0.02341	0.03570	0.03048
Rest landmarks	0.04424	0.04605	0.04608

Table 2: The average normalized distance error from the landmark detection algorithm on $D0$, $D1$, and $D1_m$. “Eye landmarks” means the landmarks of the left eye and the right eye. “Rest landmarks” indicates the landmarks of the nose and the left corner of the mouth and the right corner of the mouth.

to 0. The adoption of only one discriminator in our work is mostly for improving the computation efficiency.

6. Conclusion

In this work, we propose to tackle the task of face attribute manipulation via residual image learning which helps the image transformation networks focus on the attribute-specific area on the face image. Dual learning is also shown to be beneficial for improving the quality of generated images. Experiments demonstrate the effectiveness of the proposed method. The scheme of learning the residual data can also be promising in other relevant applications that need to model the difference between the source data and the target data.

References

- [1] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [2] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.
- [3] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.
- [4] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer. Facial attributes classification using multi-task representation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.
- [5] S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing*, 24(1):189–204, 2015.
- [6] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.
- [7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
- [8] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
- [10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
- [11] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *arXiv preprint arXiv:1603.08155*, 2016.
- [12] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.
- [13] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.
- [14] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [15] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372. IEEE, 2009.
- [16] H. Lakkaraju, E. Kamar, R. Caruana, and E. Horvitz. Discovering blind spots of predictive models: Representations and policies for guided exploration. *arXiv preprint arXiv:1610.09064*, 2016.
- [17] A. B. L. Larsen, S. K. Sønderby, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.
- [18] J. Li, J. Chen, and Z. Chi. Smile detection in the wild with hierarchical visual feature. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 639–643. IEEE, 2016.
- [19] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016.
- [20] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.
- [21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.
- [22] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, pages 52–59. Springer, 2011.
- [23] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016.
- [24] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [25] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.
- [26] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 896–903, 2013.
- [27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.
- [28] L. Theis, A. v. d. Oord, and M. Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.
- [29] W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, and N. Sebe. Recurrent face aging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [30] Y. Xia, D. He, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. Dual learning for machine translation. *arXiv preprint arXiv:1611.00179*, 2016.